the globus alliance
www.globus.org

MDS4 Deployment for TeraGrid Resource Selection

Primary Contact – Jennifer Schopf, jms@mcs.anl.gov

MCS Technical Memorandum No. ANL/MCS-TM-293

Abstract: This document describes information for using the Globus Toolkit Monitoring and Discovery System (MDS4), part of the TeraGrid r3 GT4 deployment. It includes a brief introduction to MDS4, a description of the data being collected on the TeraGrid systems, high-level deployment information, and possible next steps.

1. Project Motivation and Deliverables

One of the near-term goals for the TeraGrid project is to deploy a scheduling system that can interact with most, if not all, of the TeraGrid sites to make resource selection decisions. Progress toward this goal, however, has been impeded by the lack of a common monitoring framework across the various sites, in part due to different local policies – different sites have deployed different queuing systems (Open PBS, PBS Pro, Torque, etc.) and different cluster monitoring systems (Clumon, Nagios, Ganglia, etc.). And while similar information is available from all sites, it has not been well defined or easy to access, nor has there been a definition of the minimal set of information a site must supply to assist with resource selection.

As part of the current GT4 deployment on TeraGrid, sites will be rolling out the MDS4 infrastructure that interacts with local systems to gather data and provides a single, standard interface to the data. While there is currently no common agreement for all of the data needed to make resource selection decisions, we have defined a base set of approximately 30 attributes that will be advertised from all deploying sites, based on our previous work with scheduling systems and analysis of several common schedulers that work with queued deployments such as the TeraGrid. This data is detailed in Section 3.

As a result of the current MDS4 deployment, users will have access to a simple Web interface to examine resource selection decisions, and metaschedulers will have a common interface to the data they need across the TeraGrid, through either command line or Java APIs, as detailed in Section 3.3. As an added benefit, system administrators will be able to take advantage of the built-in trigger function and will be automatically notified of failures or conditions that change in negative ways, as described in Section 3.4.

2. MDS4 Overview

The Globus Toolkit's Monitoring and Discovery System (MDS) implements a standard Web services interface to a variety of local monitoring tools and other information sources. MDS4 is a “protocol hourglass,” depicted in Figure 1, defining standard protocols for information access and delivery and standard schemas for information representation. Below the neck of the hourglass, MDS4 interfaces to different local information sources, translating their diverse schemas into appropriate XML schema (based on standards such as the GLUE schema whenever possible). Above the neck of the hourglass, various tools and applications can be constructed to take advantage of the uniform Web services query, subscription, and notification interfaces to the information source.

MDS4 builds on query, subscription, and notification protocols and interfaces defined by the WS Resource Framework (WSRF) and WS-Notification families for specifications and implemented by the GT4 Web Services Core. Building on this base, we have implemented a range of information providers used to collect data from specific sources. These components often interface to other tools and systems.

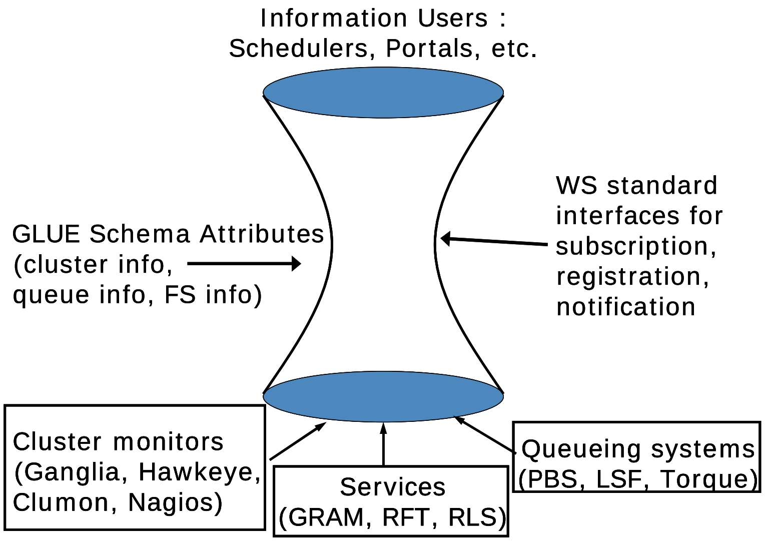


Figure 1: The MDS4 hourglass provides a uniform query, subscription, and notification interface to a wide variety of information sources, Web services, and other monitoring tools.

MDS4 also provides two higher-level services: an Index service, which collects and publishes aggregated information about information sources, and a Trigger service, which collects resource information and performs actions when certain conditions are triggered. These services are built on a common Aggregation Framework infrastructure that provides common interfaces and mechanisms for working with data sources. Additionally, a Web-based user interface called WebMDS provides a simple XSLT-transform-based visual interface to the data. Figure 2 depicts a typical MDS4 project deployment.

Additional information on MDS4 can be found online at http://www.globus.org/toolkit/mds/ and in the technical report available at http://www.mcs.anl.gov/~jms/Pubs/mds4.hpdc06.pdf .

3. MDS4 and the TeraGrid: Monitoring for Resource Selection

The TeraGrid is using MDS4 to provide a common interface to data relevant to selecting the “best” set of resources to use for a particular job, such as queue lengths and architecture types. End users (via a Web interface), metascheduling systems, and other applications will be able to use this deployment to find the resources that best meet their needs.

The TeraGrid deployment (Figure 2) consists of a set of information providers deployed at each site, as detailed in Section 3.1, an Index server at each site, a TeraGrid-wide Index (located at mds.teragrid.org), and a WebMDS server (also located at mds.teragrid.org). Table 1 lists each site, its cluster monitoring system and queuing system, and the TeraGrid release in which the information providers for these systems are available. Details on the data and deployment are given in the next sections.

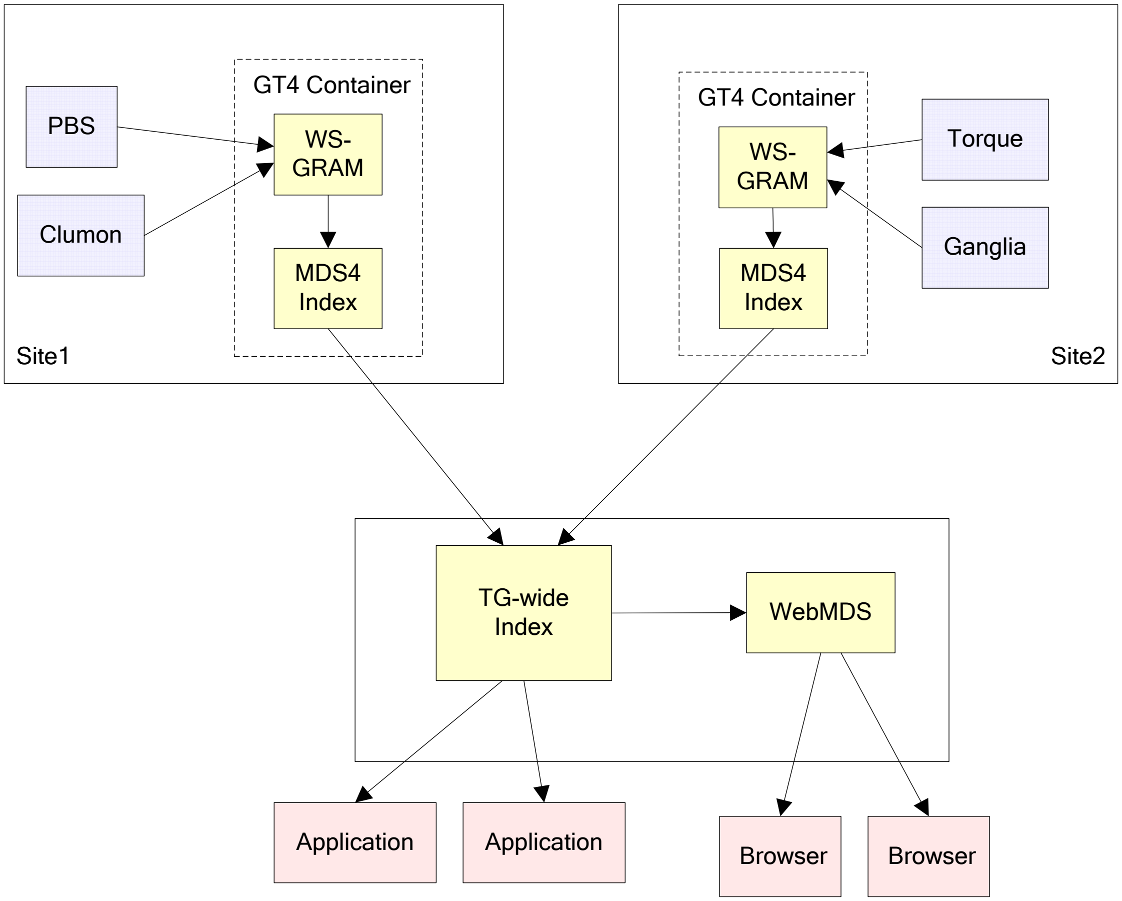


Figure 2: TeraGrid deployment of MDS4 to collect metascheduler information from various sources.

Table 1: TeraGrid sites, their queuing systems, local monitoring systems, and the release containing the needed information providers for each.

TeraGrid Machine	*Queuing System*	*IP in Release*	*Monitoring System*	*IP in Release*
iu.teragrid.org	PBSPro	R2-R3	Ganglia	R2-R3
NCSA Cobalt	PBSPro	R2-R3	Clumon	R2-R3
NCSA IA64	Torque	R2-R3	Clumon	R2-R3
ORNL	Torque	R2-R3	Nagios	R3
PSC Lemieux	PBS	R2-R3	TC SIMon?	No
PSC Rachel	PBS	R2-R3	Mon	No
PSC Red Storm	PBSPro	R2-R3	?	No
Purdue AIX	PBSPro	R2-R3	SPCiz (custom)	No
Purdue Linux	PBSPro	R2-R3	Ganglia	R2-R3
SDSC TG ia64	Torque	R2-R3	Clumon	R3
SDSC DataStar	LoadLeveler	No	IBM RSCT/ CSM	No
TACC Maverick	SGE	No	Nagios	R3
TACC Lonestar	LSF	No	Nagios	R3
uc.teragrid.org	Torque	R2-R3	Nagios, Ganglia	R3

3.1 Information Providers – Data

Information relevant to resource selection is currently available from cluster monitoring, resource management, scheduling systems, and configuration files. While there is currently no common agreement for all of the data needed to make resource selection decisions, based on our previous work with scheduling systems and analysis of several common schedulers used with queued platforms such as the TeraGrid, we have defined a set of approximately 30 attributes to gather from each site.

We currently gather queue-specific information (e.g., location of the GRAM server for a queue, the number of jobs currently waiting in that queue) and host-specific information (e.g., the CPU type and speed and the operating system). In addition, we allow for the grouping of similarly configured hosts into *clusters* and *subclusters* as a way of aggregating some of the host information. Subclusters are defined by the local site administrators such that they contain a set of hosts that have essentially the same configuration (CPU information, OS information, etc.), and data is reported for the subcluster as whole. In addition, each host record includes the name of the subcluster that the host belongs to.

Tables 2, 3, and 4 detail the data being gathered in terms of queue information, cluster information, and TeraGrid-specific data. Currently, the monitoring system includes providers that furnish queue information for PBS-compatible queueing systems, Condor, and the Globus Fork scheduler and host/cluster/subcluster information for sites running Clumon, Ganglia, Hawkeye, or Nagios. Sites running a cluster monitoring system other than these can interface to the MDS4 system by periodically creating their own .html file, which will be read in as needed. For additional information on this please contact the MDS team.

Table 2: Queue information

Attribute	Source	Dependency	Static/Dynamic
Name	PBS	GT4 GRAM	Dynamic
Unique ID	PBS	GT4 GRAM	Dynamic
GRAM version	Config file		Static
GRAM host name	Config file		Static
GRAM port/url	Config file		Static
LRMS type	PBS	GT4 GRAM	Static (dependent on IP being run)
LRMS version	PBS	GT4 GRAM	Dynamic
Total CPUs	PBS	GT4 GRAM	Dynamic
Free CPUs	PBS	GT4 GRAM	Dynamic
Queue status	PBS	GT4 GRAM	Dynamic
Total jobs	PBS	GT4 GRAM	Dynamic
Running jobs	PBS	GT4 GRAM	Dynamic
Waiting jobs	PBS	GT4 GRAM	Dynamic
Policy – max wall clock time	PBS	GT4 GRAM	Dynamic
Policy – max CPU time	PBS	GT4 GRAM	Dynamic
Policy – max total jobs	PBS	GT4 GRAM	Dynamic
Policy – max running jobs	PBS	GT4 GRAM	Dynamic

Table 3: Cluster/subcluster information

Attribute	Source	Dependency	Static/Dynamic
Type (cluster/Subcluster)	Confif file		Static
Name	Config file		Static
Unique ID	Config file		Static
Processor type	Ganglia/Clumon/Nagios		Dynamic
Processor speed	Ganglia/Clumon/Nagios		Dynamic
Total memory	Ganglia/Clumon/Nagios		Dynamic
Operating system	Ganglia/Clumon/Nagios		Dynamic
SMP size	Ganglia/Clumon/Nagios		Dynamic
Total nodes	--		Dynamic
Storage device name	Ganglia/Clumon/Nagios		Dynamic
Storage device size	Ganglia/Clumon/Nagios		Dynamic
Storage device available space	Ganglia/Clumon/Nagios		Dynamic

Table 4: Host information

Attribute	Source	Dependency	Static/Dynamic
Name	Ganglia/Clumon/Nagios		Dynamic
Unique ID	Ganglia/Clumon/Nagios + config file		Dynamic
Node properties	PBS	GT4 GRAM	Dynamic

3.2 Information Providers – Deployment

Each site that wishes to have its resources appear in the TeraGrid-wide Index server must deploy an MDS Index server for this site and perform the following:

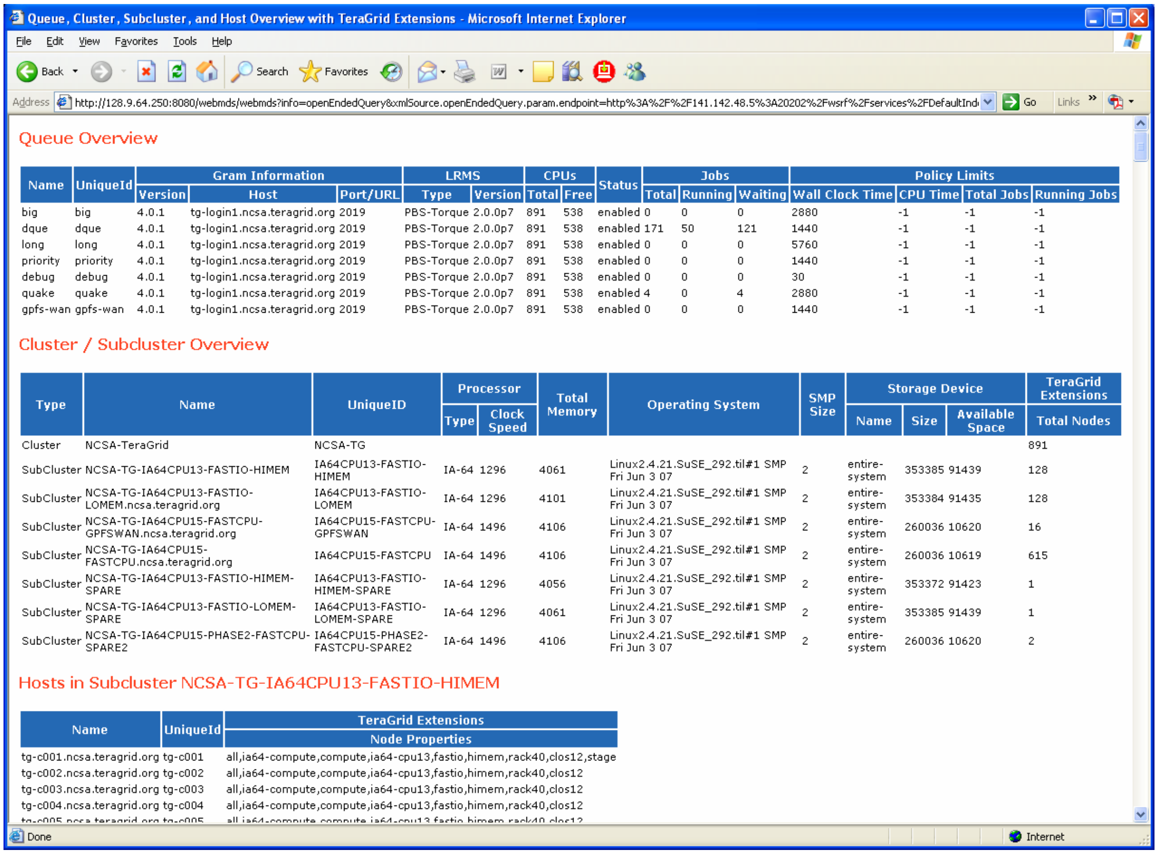
- Configure it to use the information providers appropriate for its scheduler (e.g., PBS or Condor) and cluster monitoring system (e.g., Clumon, Ganglia, or Nagios).
- Define appropriate clusters/subclusters for the site in the configuration file.
- Configure the site's Index server to feed information to the TeraGrid-wide Index server.

Details about how to perform each of these tasks for TG r2 can be found at http://software.teragrid.org/docs/ctss3/globus/4.0.1-r2/README.1st and for r3 can be found at http://software.teragrid.org/docs/ctss3/globus/4.0.1-r3/README.1st

3.3 How to Use the Information

The end user for this set of data is the Grid scheduling system that will be deployed across the TeraGrid. Four candidate systems are being considered: GRMS, Warren Smith and TACC's queue predictor system, SDSC's GAR system, and the Moab peer scheduling network. The data we are collecting will be appropriate for any of these that plan to interact with a queued system such as the TeraGrid. As the TeraGrid's metascheduling plans mature, we will work close with these groups to identify any additional requirements.

Meanwhile, we have deployed a simple Web interface using WebMDS which performs XSL transforms that allow users to easily see the full set of data needed to make their own resource selection decisions, as shown in Figure 3. This data can also be retrieved directly from the TeraGrid-wide Index server by using Java or command-line calls, as detailed in the GT 4.0 Index Service Users Guide (http://www.globus.org/toolkit/docs/4.0/info/index/user-index.html).


Queue, Cluster, Subcluster, and Host Overview with TeraGrid Extensions - Microsoft Internet Explorer
File
Edit
View
Favorites
Tools
Help
Back
Search
Favorites
Address
http://128.9.64.250:8080/webmds/webmds?info=openEndedQuery&xmlSource.openEndedQuery.param.endpoint=http%3A%2F%2F141.142.48.5%3A20202%2Fwsrf%2Fservices%2FDefaultInd
Go
Links
Done
Internet

Queue Overview

Name	UniqueId	Gram Information			LRMS		CPUs		Status	Jobs			Policy Limits			
		Version	Host	Port/URL	Type	Version	Total	Free		Total	Running	Waiting	Wall Clock Time	CPU Time	Total Jobs	Running Jobs
big	big	4.0.1	tg-login1.ncsa.teragrid.org	2019	PBS-Torque	2.0.0p7	891	538	enabled	0	0	0	2880	-1	-1	-1
dque	dque	4.0.1	tg-login1.ncsa.teragrid.org	2019	PBS-Torque	2.0.0p7	891	538	enabled	171	50	121	1440	-1	-1	-1
long	long	4.0.1	tg-login1.ncsa.teragrid.org	2019	PBS-Torque	2.0.0p7	891	538	enabled	0	0	0	5760	-1	-1	-1
priority	priority	4.0.1	tg-login1.ncsa.teragrid.org	2019	PBS-Torque	2.0.0p7	891	538	enabled	0	0	0	1440	-1	-1	-1
debug	debug	4.0.1	tg-login1.ncsa.teragrid.org	2019	PBS-Torque	2.0.0p7	891	538	enabled	0	0	0	30	-1	-1	-1
quake	quake	4.0.1	tg-login1.ncsa.teragrid.org	2019	PBS-Torque	2.0.0p7	891	538	enabled	4	0	4	2880	-1	-1	-1
gpfs-wan	gpfs-wan	4.0.1	tg-login1.ncsa.teragrid.org	2019	PBS-Torque	2.0.0p7	891	538	enabled	0	0	0	1440	-1	-1	-1

Cluster / Subcluster Overview

Type	Name	UniqueID	Processor		Total Memory	Operating System	SMP Size	Storage Device			TeraGrid Extensions
			Type	Clock Speed				Name	Size	Available Space	Total Nodes
Cluster	NCSA-TeraGrid	NCSA-TG									891
SubCluster	NCSA-TG-IA64CPU13-FASTIO-HIMEM	IA64CPU13-FASTIO-HIMEM	IA-64	1296	4061	Linux2.4.21.SuSE_292.til#1 SMP Fri Jun 3 07	2	entire-system	353385	91439	128
SubCluster	NCSA-TG-IA64CPU13-FASTIO-LOMEM.ncsa.teragrid.org	IA64CPU13-FASTIO-LOMEM	IA-64	1296	4101	Linux2.4.21.SuSE_292.til#1 SMP Fri Jun 3 07	2	entire-system	353384	91435	128
SubCluster	NCSA-TG-IA64CPU15-FASTCPU-GPFSWAN.ncsa.teragrid.org	IA64CPU15-FASTCPU-GPFSWAN	IA-64	1496	4106	Linux2.4.21.SuSE_292.til#1 SMP Fri Jun 3 07	2	entire-system	260036	10620	16
SubCluster	NCSA-TG-IA64CPU15-FASTCPU.ncsa.teragrid.org	IA64CPU15-FASTCPU	IA-64	1496	4106	Linux2.4.21.SuSE_292.til#1 SMP Fri Jun 3 07	2	entire-system	260036	10619	615
SubCluster	NCSA-TG-IA64CPU13-FASTIO-HIMEM-SPARE	IA64CPU13-FASTIO-HIMEM-SPARE	IA-64	1296	4056	Linux2.4.21.SuSE_292.til#1 SMP Fri Jun 3 07	2	entire-system	353372	91423	1
SubCluster	NCSA-TG-IA64CPU13-FASTIO-LOMEM-SPARE	IA64CPU13-FASTIO-LOMEM-SPARE	IA-64	1296	4061	Linux2.4.21.SuSE_292.til#1 SMP Fri Jun 3 07	2	entire-system	353385	91439	1
SubCluster	NCSA-TG-IA64CPU15-PHASE2-FASTCPU-SPARE2	IA64CPU15-PHASE2-FASTCPU-SPARE2	IA-64	1496	4106	Linux2.4.21.SuSE_292.til#1 SMP Fri Jun 3 07	2	entire-system	260036	10620	2

Hosts in Subcluster NCSA-TG-IA64CPU13-FASTIO-HIMEM

Name	UniqueId	TeraGrid Extensions
		Node Properties
tg-c001.ncsa.teragrid.org	tg-c001	all,ia64-compute,compute,ia64-cpu13,fastio,himem,rack40,clos12,stage
tg-c002.ncsa.teragrid.org	tg-c002	all,ia64-compute,compute,ia64-cpu13,fastio,himem,rack40,clos12
tg-c003.ncsa.teragrid.org	tg-c003	all,ia64-compute,compute,ia64-cpu13,fastio,himem,rack40,clos12
tg-c004.ncsa.teragrid.org	tg-c004	all,ia64-compute,compute,ia64-cpu13,fastio,himem,rack40,clos12
tg-c005.ncsa.teragrid.org	tg-c005	all,ia64-compute,compute,ia64-cpu13,fastio,himem,rack40,clos12

Figure 3: WebMDS for TeraGrid.

3.4 *Trigger Service*

The MDS-Trigger service collects information and compares that data against a set of conditions defined in a configuration file. When a condition is met, an action takes place, such as emailing a system administrator when the disk space on a server reaches a threshold.

TeraGrid sites can set up this service so that they are alerted when a monitoring system or queueing system goes down, when queue lengths get overly long, or when the space on a storage system exceeds a set limit. Information on how to set up this service can be found at http://www.globus.org/toolkit/docs/4.0/info/trigger/index.html

4. Next Steps

The current MDS4 deployment for TeraGrid meets the basic requirements for information provided to allow resource selection. Two pieces of cluster data that have been requested in addition are a node to queue mapping, which would be done through a static configuration file since this data cannot be discovered dynamically programmatically, and a listing of the softenv keys for a cluster. Delivery dates for these pieces of information have not yet been set. In addition, we have been in discussion with users to provide a simple resource selection interface using a web browser and pull-down menus for basic resource attributes, and then showing the resulting queue lengths for the resources that match.

Additional data and interfaces can be provided as needed, and we expect this deployment to grow as it receives more use and experience.

5. Additional Information

Additional information on MDS4 can be found online at http://www.globus.org/toolkit/mds/ and in the technical report available at http://www.mcs.anl.gov/~jms/Pubs/mds4.hpdc06.pdf .

TG Deployment Documentation r2 http://software.teragrid.org/docs/ctss3/globus/4.0.1-r2/README.1st
and for r3 at http://software.teragrid.org/docs/ctss3/globus/4.0.1-r3/README.1st

Trigger Service http://www.globus.org/toolkit/docs/4.0/info/trigger/index.html

Ganglia: http://ganglia.sourceforge.net/

Clumon: http://clumon.ncsa.uiuc.edu/

Acknowledgments

This work was supported in part by the Mathematical, Information, and Computational Sciences Division subprogram of the Office of Advanced Scientific Computing Research, Office of Science, U.S. Department of Energy, under contract W-31-109-Eng-38. Additional support was provided by NSF NMI Award SCI-0438372.